Signal-driven sound processing for uncontrolled environments

Krijnders, Johannes Dirk

2

Basic signal processing

2.1 Design criteria

Most applications mentioned in section 1.1 require a system that must work out-of-the-box, i.e. without the intervention of a specialist. The system should work in cities, villages, countryside and should be undisturbed by influences of the weather or other changes in the environment. For most applications it is not necessary to recognize all sound sources perfectly, but the general pattern of sound sources should provide a reliable indication of the sonic environment. These requirements pose demands on the design of the system:
The system should be able to

- handle concurrent sources (section 2.1.2)
- deal with changing transmission effects (section 2.1.3)
- handle diverse, possible unknown, classes (section 2.1.4)

The next section will discuss what a (environmental) sound source is and the following sections describe the problems associated with the demands mentioned above.

2.1.1 Sound sources

Based on the dictionary (New Oxford American Dictionary) definitions of "sound" and "source" a sound source would be "a place, person, or thing from which something comes", where something is "vibrations that travel through the air or another medium and can be heard when they reach a person's or animal's ear". This would entail that a sound source is the place where the vibrations originate, but this "place" is unclear when considering the common sense definition that for example a car is a sound source, while the sound source would technically be the explosions in the engine, the air flow around the car and the noise from the interaction between the tires and the road. The sound source is thus, like the auditory objects in section 1.3, dependent on the detail level the user is interested in.

In some literature a further qualifications like "environmental" (Cowling and Sitte, 2003; Shafiro and Gygi, 2004), or "everyday" are given to a sound source. But clear definitions of these qualifications are lacking. It is usually assumed that environmental sounds exclude speech or music, however these terms are not clearly defined either. Ballas and Howard (1987) and Vanderveer (1979) agree on two criteria for environmental sounds: First they are "produced by real events" and second they convey "meaning by virtue of the causal events".

2.1.2 Concurrent sound sources

As sound sources seldom occur in isolation in the real world it is important to handle concurrent sources. Usually this problem is treated as a sound source with added noise (Hayes, 1996, Chapter 7):

$$x(t) = x_{target}(t) + n(t) \tag{2.1}$$

$$n(t) \sim \mathcal{N}(\mu, \sigma^2) \tag{2.2}$$

This approach assumes that the non-target sources add up to normally distributed noise, which is only true for certain (noisy) sources or when many uncorrelated sources are present (central limit theorem). The assumption can be valid in cases where there is much control over the recording and the main noise source is known, for example a telephone with close talking microphone increases the probability of a single source and the noise may be caused by electrical and thermal sources which can be assumed to be broadband noise. The more general problem can be stated as (Cardoso and Martin, 2007):

$$x(t) = \sum_{n=1}^{N} a_n x_{s,n}(t) \tag{2.3}$$

where N is the number of sound sources, a_n represents the sound level decrease due to the distance between the sound source and the receiver and

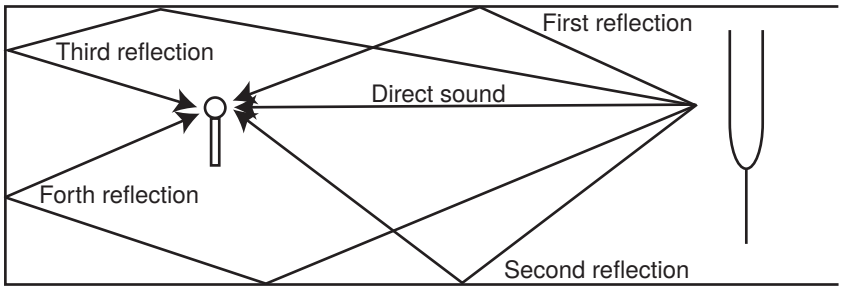


Figure 2.1: The path of the direct sound and the first four reflections in a room.

$x_{s,n}(t)$ is the time signal of an individual source at that source. This problem is a standard inverse problem and cannot be solved without extra knowledge if the number of microphones (observations of $x(t)$) is less then the number of sources (N), i.e. the system is underdetermined. Hence, most current approaches use multiple microphones and independent component analysis (ICA, (Choi et al., 2005)) or beam-forming techniques(Kellerman, 2009) to extract to separate $x_n(t)$. This changes equation 2.3 to

$$x_m(t) = \sum_{n=1}^{N} a_{n,m} x_{s,n}(t) \tag{2.4}$$

$$\mathbf{x}_m = A\mathbf{x}_s \tag{2.5}$$

where $A = a_{n,m}$ and the matrix formulation is the standard problem statement of ICA. And this problem is solvable when the matrix A is constant (or slowly changing) and only one of the sound sources $\mathbf{x}_s$ produces broadband noise. This entails that the range of application is limited to locations where the acoustics are (quasi-)constant and the sources are (quasi-)stationary. This prevents ICA from being applicable in the applications mentioned in section 1.1.

2.1.3 Transmission effects

As sound travels from the source to the microphone it is not only attenuated and mixed with other sources, the sound also reflects off surfaces leading the multiple paths from source to microphone 2.1. The indirect paths are longer and thus a delayed and more attenuated version of the original sound arrives at the microphone. Apart from being delayed, the frequency content of the sound changes as surfaces don't necessarily reflect all frequencies equally well. This process is called reverberation.

Reverberation is usually modeled as an impulse response. It can be obtained by recording of the sound of a perfect impact or a swept sine (Farina, 2005), or be derived by modeling the room and the object inside it. The sound

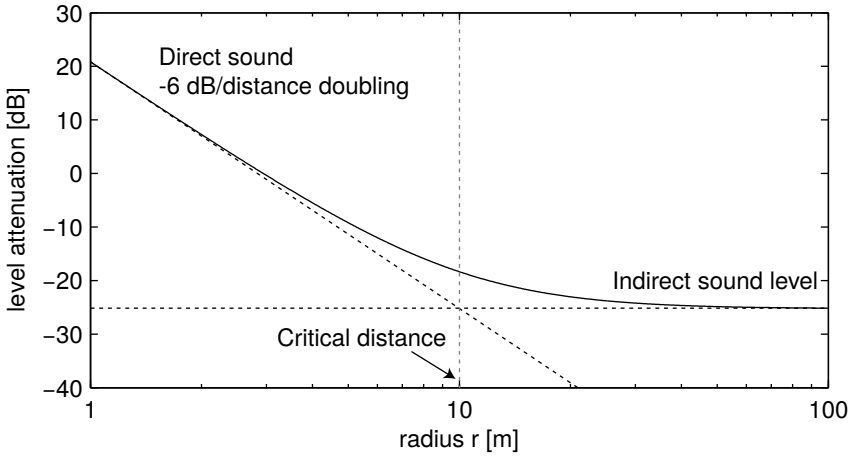


Figure 2.2: The attenuation of the sound pressure level in a ideal room of $50m^3$. The critical distance is the distance where the level of the indirect sound is equal to the direct sound, in this case $10m$.

arriving at the microphone can be simulated by convolving the source sound with the impulse response, thus replacing the simple factor in equation 2.3 with a convolution:

$$x_m(t) \quad = \quad \mathbf{r} \otimes x_s(t) \tag{2.6}$$

The impulse ($\mathbf{r}$) is dependent on both the position of the source and the receiver and on the properties of the room and objects in it. If source and receiver are close in comparison to the size of the room the direct sound contributes most energy to the receiver. As source and receiver move away from each other the relative contribution of the direct sound compared to the indirect sound (constant) decreases (see figure 2.2). The distance from the source at which the contribution for the indirect sound equals the contribution from the direct sound is called the critical distance. Within this ratio speech is understandable independent of the amount of reverberation, outside the intelligibility is a function of the amount of reverberation (Peutz, 1971). Our system will need to work both inside (Chapter 8) and outside (Chapter 6 and 7) the reverberation radius.

Reverberation complicates the mixing of sources by adding time-delayed, frequency-dependently attenuated copies each source. The only interval that the sound is undisturbed by delayed copies is when the direct sound has arrived but first reflection has not. These intervals are very short for rooms (3 ms = 1 meter path length difference). Yet this property is exploited by humans for source localization and source identification. This exploitation

is called the precedence effect (Litovsky et al., 1999). To exploit this effect the system needs to detect the unperturbed start of the sound, however current systems are not build, nor suitable, to do that.

Current techniques to deal with reverberation can be split in two categories (Habets, 2007): reverberation suppression and reverberation cancelation. This last category requires a full estimate of the impulse response and is therefore unsuitable for changing environments (Haykin, 1994, 2000). Reverberation suppression requires less knowledge of the transmission properties, but needs more information about the sources present as these methods use knowledge of the source. This knowledge can include the kind of source (Deller Jr. et al., 1999), e.g. speech for linear prediction coding, or location for beam-forming techniques (Trees, 2002). Since the sources in an open environment are unknown, this requirement is a problem.

2.1.4 Unknown sources

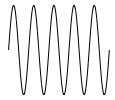
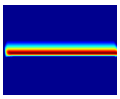
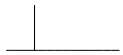
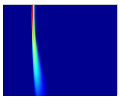
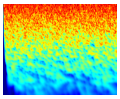
When the system is deployed in an uncontrolled environment it will encounter sound sources that it has no knowledge of. It will not be able to classify these sources, but it would be beneficial if the recordings of these sources are stored for further analysis. Current systems expect that all encountered classes were part of their training database. This entails that they assign a class to every part of the signal they try to classify, regardless of whether that class actually matches the signal, but it just happened to have a slightly higher probability then the other classes. The only mechanism those systems have to ignore unknown sources is to ignore sources that do not exceed a certain threshold of probability.

2.2 On tones and pulses

To extract evidence that is likely to stem from a single source, we need a criterium that allows us to do so, solely based on the signal at hand. Sound production can roughly be divided into three different mechanisms: resonance, impact and turbulence (Gaver, 1993b). These mechanisms result in respectively tones, pulses and broadband noise[1]. The first two are very localized, tones in frequency and pulses in time. This makes that the chance for two tones or pulses to mask each other in the time-frequency plane is small. As such tones and pulses that stem from a single source have a large chance to be found as a single time-frequency region.

[1]The term “noise” will be used in this thesis in the meaning of colored or white noise, i.e. noise resulting from aperiodic processes. Compare to “ruis” in dutch or “Krach” in german (Dubois and Guastavino, 2008)

Table 2.1: Limits of the Heisenberg inequality

Tones	$\lim_{\sigma_t \to 0} \sigma_t \sigma_\omega = \frac{1}{2}$		
Pulse	$\lim_{\sigma_\omega \to 0} \sigma_t \sigma_\omega = \frac{1}{2}$		
Broadband	$\lim_{\sigma_t \sigma_\omega \to \infty}$		

The three mechanisms match the extremes of the Heisenberg inequality (see table 2.1). The Heisenberg inequality states that:

$$\sigma_t \sigma_\omega \quad \geq \quad \frac{1}{2} \tag{2.7}$$

where σ_t is the time variation and σ_ω is the frequency variation. This notation follows Hut et al. (2006), but the inequality was first derived by Gabor (1946). Gröchenig (2001, Chapter 2) discusses the uncertainty relation extensively.

The localization of tones and pulses allows for tracking them through time, resp. frequency. This tracking provides extra certainty and groups together similar parts of the spectrum.

2.3 Cochleogram

As we are interested in tones and pulses and we want to track them we need a continuous development of tones and pulses through our time-frequency representation without noticeable biases for special frequencies or points in time. Time-frequency representations decompose the audio signal into two-dimensional matrices where time is one dimension and frequency the other. The most popular one is the spectrogram based on the short-term fast Fourier transform (SFFT). However this Fourier transform does have biases, both in time and frequency, see figure 3.13 . An alternative is to use a basilar membrane or cochlea model which does not have these biases and thus makes it easier to track continuous signals through time and frequency.

Figure 2.3: Location of the cochlea (left, purple) and a cross-section of the cochlea (right) along the gray line. The location of the basilar membrane is highlighted in red . The membrane vibrates when sound reaches the ear and the haircells on the membrane translate the amplitude change to nerve signals.

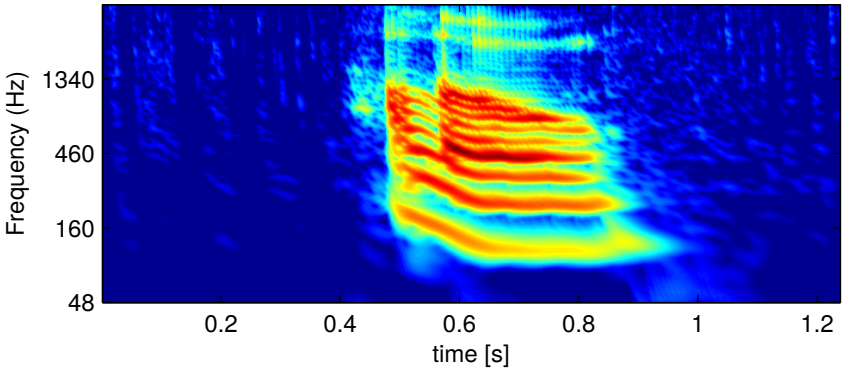


Figure 2.4: Cochleogram of the author saying “hallo”, the noise is from surroundings

2.3.1 Basilar membrane

The basilar membrane is the last mechanical part of the human auditory system. Its movement triggers responses in the haircells on its surface. The basilar membrane divides the cochlear duct in two halves (figure 2.3). This cochlear duct is rolled in a snail-like shape and reduces in diameter from the entrance to the end. This reduction changes the frequency properties of the basilar membrane and thus different frequencies in the sound stimulate different haircells.

Many models of this system have been proposed, either with transmission-line models which model the physics of the cochlea (Duifhuis et al., 1985) or with filterbank models (Irino and Patterson, 1997) which try to match the psycho-acoustical data.

The transmission-line model has shown to be more accurate in resolving sound close to the Heisenberg limit (eq. 2.7) than the gamma-tone filterbank Hut et al. (2006). So for our purposes the transmission-line model would be better, however due to the wide-spread use of the gamma-tone filterbank we use the gamma-tone filterbank through out this thesis. Most experiments have also been done with the Duifhuis et al. (1985) model and the results were very similar.

The gamma-tone or gamma-chirp filterbank (Irino and Patterson, 1997) is a widely used model for the basilar membrane. Its filter coefficients (h_{gc}) are defined by ($c = 0$ for the gamma-tone):

$$h_{gc} = at^{N-1}e^{-2\pi bB(f_c)t}e^{j(2\pi f_c t + c\log(t))} \tag{2.8}$$

where f_c is the center frequency of the channel, N the order of the gamma-tone (4) and $a = 1$, $b = 0.71$ and $c = -3.7$. These values are somewhat different than in Irino and Patterson (1997) in favor of a narrow tonal response (at the cost of increased group delay) to make its response closer to that of the transmission-line model. The frequency range extends from 60 Hz to 4000 Hz. The center frequencies of the filterbank are distributed logarithmically. The bandwidth of each filter is given by the ERB scale (Moore and Glasberg, 1996):

$$B(f_c) = 24.7 + 0.108f_c \tag{2.9}$$

2.3.2 Cochleogram

The basilar membrane models result in a amplitude representation of the membrane, this gives very detailed information on the signal which may be useful (Krijnders et al., 2007). However for our purposes a energy representation is more convenient. To calculate the energy in the amplitude matrix the filter output is squared and leaky integrated with a channel dependent time-constant $\tau_c = \max(5, 2/f_c)$ ms. This leaky-integration method yields, in

combination with the logarithmic frequency axis, a constant-Q-like representation . The filterbank output is squared to represent an energy measure, down-sampled to 200 Hz, and compressed logarithmically to express the energy in dB. The resulting representation is a spectrogram-like representation, termed a cochleogram, with 5 ms frames. For a full mathematical description see appendix A.

Both the frequency and the energy representation are logarithmic and comply with Weber's law, which entails that they are able to represent many orders of magnitude in a limited dynamic range. Both are central properties of auditory processing. For historical reasons "channels" are often referred to as "segments" and we will use both terms.

2.4 Local target to non-target ratio

In speech recognition research the signal to noise ratio (SNR) is often used as a measure of how bad a signal is degraded by mixing with other sources (noise) than the target signal. It is defined as the ratio of the power of the signal to the power of the noise:

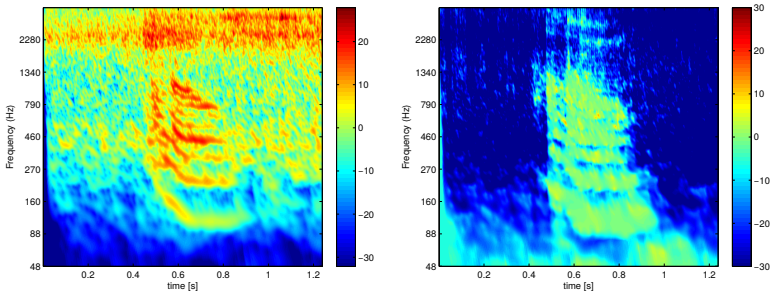
$$\text{SNR} = 10\log\left(\frac{P_{signal}}{P_{disturbance}}\right) \tag{2.10}$$

This ratio is calculated for a complete speech sample and thus a global measure. However spectral and temporal content of the signal and the noise are often different, so the SNR says nothing about the ratio of power at specific moments in time and frequency. This problem is sometimes alleviated by using A-weighting (S1.4, 2001, based on Fletcher and Munson, 1933) both the speech and noise signal and using speech detection (P.56, 1993). This last method prevents the silences in speech from changing the SNR values.

From speech research started by Fletcher (Fletcher, 1950) and continued by others (Allen, 1994; Bronkhorst, 2000) it is known that only the local target to non-target ratio counts in human performance. Cooke (2006) proposed a model for speech recognition that uses this fact under the term glimpsing at the target when the LSNR ratio is advantageous to do so.

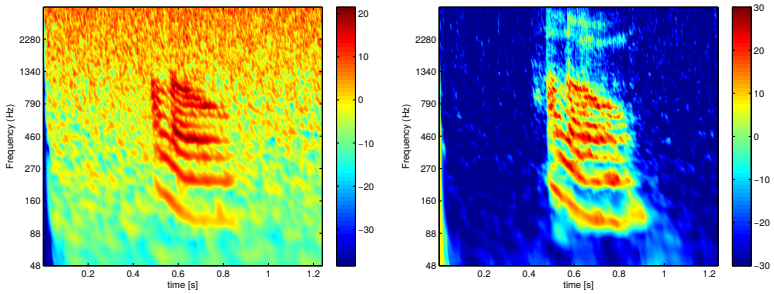
Moreover, and this is more a naming question, it assumes that signal and noise are well defined, but if two people talk at the same time both may be considered signal. For this reason the term target-to-non-target ratio is more appropriate.

Figure 2.5 shows how different kinds of noise influence the LSNR ratio. The first column shows the energy distributions, the second the LSNR distribution. Calculating the SNR using speech detection and A-weighting results, as expected, in lower LSNR values compared to the global SNR calculations



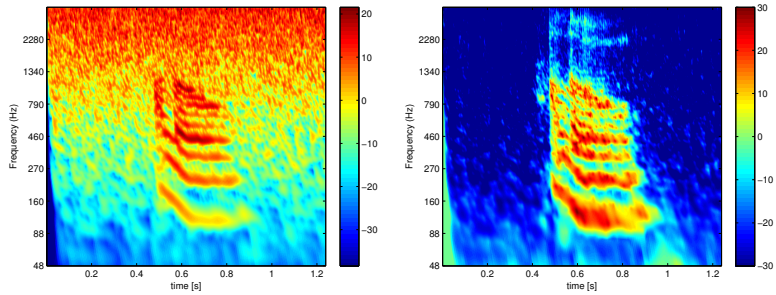
(a) Signal from figure 2.4 with subway noise added at zero dB SNR following A-weighting and ITU-T P.56 speech detection

(b) Local target to non target ratio of the signal in figure 2.5(a)



(c) Signal from figure 2.4 with pink noise added at zero dB

(d) Local target to non target ratio of the signal in figure 2.5(c)



(e) Signal from figure 2.4 with white noise added at zero dB

(f) Local target to non target ratio of the signal in figure 2.5(e)

Figure 2.5: The left column shows the clean “hallo” from figure 2.4 with several types of noise added at 0 dB SNR and the right column the local target to non target ratio of the sounds

in figures 2.5(c-f).

$$\mathrm{LSNR}(t,f) \quad = \quad 10\log\left(\frac{P_t(t,f)}{\sum_{n\neq t} P_n(t,f)}\right) \tag{2.11}$$

where P_t is the energy of the target sound and P_n is the power of all individual sources.

2.5 Scope of this thesis

Based the design criteria (section 2.1) we formulate the goals of the work presented in this thesis.

Select regions from a cochleogram that are highly likely to stem from a single source in realistic acoustical circumstances

Chapter 3 introduces a new method for signal processing. It exploits the properties of tones and pulses to extract segments in the cochleogram that are likely to belong to a single source.

Show a possible recognition strategy for these regions

Chapter 4 shows a recognition strategy based on forming groups of regions found with the methods in chapter 3. Nearest-neighbor algorithms are used to classify these groups. Initially it is not required to be perfectly correct. However, the general pattern of sources must provide a good indication of the actual auditory scene, wherever the system is deployed.

Show integration with knowledge driven context information

As the contents of the datasets is highly ambiguous, the performance of signal-driven techniques will not be maximally high. To disambiguate some detected sound event it is coupled with a knowledge-driven network. This combination is described in chapter 4

Show performance on several different datasets

Before performance on any dataset can be shown, it is necessary to create a ground truth for the recordings being analyzed. Chapter 5 discusses this process and its problems for environmental sounds.

Part III of this thesis shows the performance of the recognition system on diverse datasets and tasks. Chapter 8 extends the grouping with a formant detection algorithm and results of vowel identification are shown. Chapter 6 shows the application to recordings from a train station with a focus on the

detection of verbal aggression and related classes. Also the integration with results from video processing is discussed. Finally in chapter 7 the methods are applied to a dataset of city sounds and combined with a dynamic network to supply context information. This addition improves performance. The dataset used here is a large dataset of recording from the town of Assen. The number of classes in this dataset is high (N=54).